Australian Government
Department of Defence
Defence Science and
Technology Organisation

Review of Literature on Probability of Detection for Liquid Penetrant Nondestructive Testing

C.A. Harding and G.R. Hugo

Maritime Platforms Division
Defence Science and Technology Organisation

DSTO-TR-2623

ABSTRACT

A review of the published literature on the reliability of liquid penetrant testing (LPT) identified twelve major probability of detection (POD) studies conducted between 1968 and 2009. Based on these studies, significant variability in performance is inferred between different implementations of the post-emulsifiable LPT process. This report presents statistical inferences for the defect size expected to be detected with 90% POD by most implementations of post-emulsifiable LPT, based on the published data. The reliably-detectable defect size of a_{NDI} = 3 mm currently specified for the Royal Australian Air Force general procedure for post-emulsifiable LPT is consistent with estimates of the average performance of LPT demonstrated in the literature.

Published by

Maritime Platforms Division
DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia

Telephone: (03) 9626 7000
Fax: (03) 9626 7999

AR-015-142
November 2011

APPROVED FOR PUBLIC RELEASE

Review of Literature on Probability of Detection for Liquid Penetrant Nondestructive Testing

Executive Summary

Appropriate application of nondestructive testing (NDT) methods is dependent on knowledge of the minimum sizes of defects that the techniques are capable of reliably detecting, relative to the defect sizes that could be structurally significant. For some applications, the failure of NDT to detect a single defect could cause catastrophic failure including loss of life. Quantitative assessment of the reliability of NDT in terms of probability of detection (POD) is an essential part of aircraft structural integrity management. This report is the first in a series of literature reviews being undertaken by DSTO to examine the probability of detection for standard nondestructive testing methods used on ADF aircraft.

A review of the published literature on the reliability of liquid penetrant testing (LPT) identified twelve major POD studies conducted between 1968 and 2009. The available data from these activities have been compiled into a summary of detectable defect sizes demonstrated for LPT and the findings compared with the existing standard limitations used by the Royal Australian Air Force. All available POD data relate to post- emulsifiable LPT; no data were found in the literature that specifically address reliability of solvent-removable (method C) penetrant inspections.

The published data indicate significant variability in the performance of liquid penetrant testing between different implementations of the post-emulsifiable LPT process. The minimum reliably detectable defect size of $a_{NDI} = 3.0$ mm currently assumed by the RAAF for LPT is consistent with estimates of the *average* performance of LPT demonstrated in the literature. Consequently, the available data do not support any reduction in the existing a_{NDI} for LPT. Based on the available published data, the smallest a_{NDI} that might reasonably be predicted to be detected with 90% probability of detection by *most* implementations of post-emulsifiable liquid penetrant testing processes is 5 mm.

It is noted that the detectable defect sizes recommended in the USA Joint Service Specification Guide Aircraft Structures, JSSG-2006, for damage tolerance analysis of aircraft structures are consistent with the defect sizes demonstrated by POD trials to be reliably detected by LPT. However, the detectable defect sizes recommended in MIL-HDBK-1783B for automated fluorescent penetrant inspections of engine components are much smaller than those reported in the published POD data for LPT.

This page is intentionally blank

Authors

Dr Cayt A Harding
Maritime Platforms Division

Dr Cayt Harding has worked in the nondestructive evaluation (NDE) group of the Defence Science and Technology Organisation since 1999. She completed her undergraduate training at the University of Melbourne, receiving first class honours in Mechanical Engineering and a science degree in Applied Mathematics and Physics. Cayt has twelve years experience in research on reliability assessment of nondestructive evaluation and received a PhD from The University of Melbourne for her research on probability of detection measurement and modelling.

Dr Geoffrey R Hugo
Maritime Platforms Division

Dr Geoff Hugo is the Science Team Leader for NDE research for Air vehicle applications. He has Bachelor of Science and Bachelor of Engineering (with Honours) degrees in Applied Mathematics and Materials Engineering and a PhD in Materials Engineering, all from Monash University. Since joining DSTO in 1986, he has worked on a variety of projects in materials engineering and nondestructive evaluation. He has fifteen years experience in research on advanced computer-based non-destructive evaluation systems, including both ultrasonic and eddy-current systems. He is the author of more than 25 scientific papers and technical reports and is a member of the Institution of Engineers, Australia.

This page is intentionally blank

Contents

This page is intentionally blank

Abbreviations

AFHR	aircraft flying hours
ASNT	American Society for Nondestructive Testing
ASTM	American Society for Testing and Materials
CASR	Center for Aviation Systems Reliability
FAA	Federal Aviation Administration
FPI	fluorescent penetrant inspection
FPT	fluorescent penetrant testing
IR	infrared
LPI	liquid penetrant inspection
LPT	liquid penetrant testing
MLE	maximum likelihood estimation
NATO	North Atlantic Treaty Organisation
NDE	nondestructive evaluation
NDI	nondestructive inspection
NDT	nondestructive testing
NDTSL	Nondestructive Testing Standards Laboratory, RAAF
NRC IAR	National Research Council Institute for Aerospace Research (Canada)
POD	probability of detection
RAAF	Royal Australian Air Force
USAF	United States Air Force

Symbols

a	defect size
a_{crit}	critical defect size
a_{NDI}	minimum reliably detectable defect size
a_{90}	defect size having 90% probability of detection
$a_{90/95}$	defect size having 90% probability of detection demonstrated with 95% statistical confidence
Ψ	probability of detection
$\hat{\Psi}$	estimated probability of detection
Ψ_L	lower confidence limit on probability of detection
n	sample size
t	thickness

This page is intentionally blank

1. Introduction

Liquid penetrant testing (LPT) is a non-destructive testing[1] (NDT) method used extensively in the aerospace industry to detect surface breaking cracks in metal components. It is relied on to assure the structural integrity of aircraft through inspections of critical components during production and throughout the aircraft service life. Failure of LPT to detect defects could have a variety of consequences including unavailability of aircraft, increased maintenance costs, or catastrophic failure of safety-critical structure.

Knowledge of the reliability achieved by NDT methods, including LPT, forms an essential part of aircraft structural integrity management. In particular, the minimum reliably detectable defect size, a_{NDI}, is key information that is required as input for damage tolerance analyses, which are used to determine safe inspection intervals. The detectable defect size, and the reliability with which it can be detected, are dependent on many factors, including the inherent variability in the characteristics of the defects to be detected.

The reliability of NDT is commonly characterised in terms of the probability of detection (POD, Ψ) of a specified type of defect as a function of defect size, a. Probability of detection is traditionally determined through large-scale trials of NDT procedures on representative components to gather data for statistical analysis, which can be prohibitively expensive. To account for sampling variability inherent in any empirical statistical trial, it is normal to apply confidence limits to the estimated POD. Figure 1 shows a typical estimated POD curve, $\hat{\Psi}$, and lower confidence limit, Ψ_L, where Ψ_L

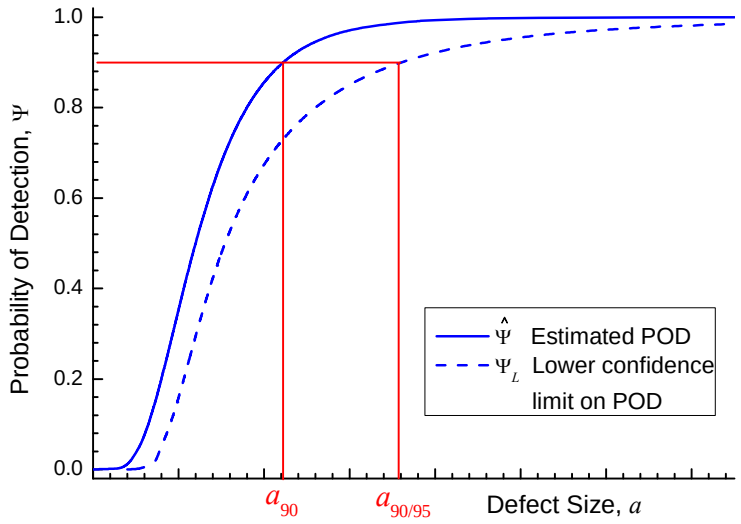

Figure 1 Estimated probability of detection, $\hat{\Psi}$, and lower confidence limit on POD, Ψ_L, plotted against defect size, a, showing the definition of a_{90} and $a_{90/95}$ values

[1] Also known as nondestructive inspection (NDI) and nondestructive evaluation (NDE). These terms are regarded as synonymous for the purposes of this report.

represents the lower bound on where the true POD curve might lie and still be consistent with the observed data [1]. Two defect sizes are frequently extracted from POD information:

a_{90} is the defect size at which the estimated POD, $\hat{\Psi}$, reaches 90%, i.e. $\hat{\Psi}(a_{90}) = 0.9$, and

$a_{90/95}$ is the defect size at which the lower 95% confidence limit Ψ_L reaches 90% POD, i.e. $\Psi_L(a_{90/95}) = 0.9$.

For aircraft with an airworthiness certification based on safety-by-inspection, a 'safe' inspection interval is determined as a prescribed fraction (typically half) of the time in aircraft flying hours (AFHR) it takes for an assumed defect to grow from the minimum detectable defect size, a_{NDI}, to the critical defect size, a_{crit}, at which the structure could fail under service loads, Figure 2. Airworthiness standards specify the defect size that is appropriate for use as a_{NDI}. For example, JSSG-2006 *Joint Service Specification Guide Aircraft Structures* is the multi-service guide to the specification of Aircraft Structures for use within the USA Department of Defence [2]. Under JSSG-2006, the recommended value for a_{NDI} is the defect size for which a 90% probability of detection has been demonstrated with 95% statistical confidence, denoted $a_{90/95}$ as above. This is the default standard for all damage tolerance analyses of airframe structure for US-built military aircraft. Other relevant airworthiness standards include the US Department of Defense Handbook for *Engine Structural Integrity Program*, MIL-HDBK-1783B [3], and the UK Ministry of Defence Standard, DEF STAN 00-970 *Design and Airworthiness Requirements for Service Aircraft*. The requirements of these standards are reviewed in greater detail in reference [1]. The default minimum reliably detectable defect sizes given in these standards are presented in Section 4, where they are compared to the findings from the literature review for reliability of liquid penetrant inspection.

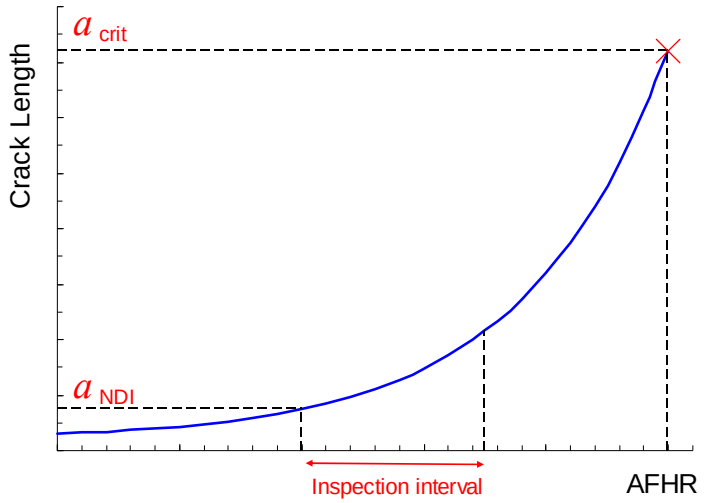

Figure 2 Inspection interval determined from a_{NDI} and crack growth curve

In practice, it is relatively rare for the a_{NDI} values used in a damage tolerance analysis to be directly underpinned by POD data from an experimental POD trial. The more common approach is to base the a_{NDI} used for damage tolerance analysis on an estimated 'limitation' for the technique, which is the smallest defect that a published NDT procedure is expected to reliably find. [2] In RAAF practice:

> *"NDT Procedure limitations state the type and size of the defect the procedure will readily detect. Limitations are intended only as a guide to engineering staff to assist in the determination of test intervals or the safe working life of an item."* (AAP 7001.068(AM1) paragraph 20) [4]

The limitation is determined based either on laboratory experiments applying the technique to simulated defects (such as machined notches) or, more frequently, from previously accepted values for similar inspection procedures and previous experience with the NDT technique.

This report is the first in a series of reviews being undertaken by DSTO to address the probability of detection for standard nondestructive testing methods. The findings from the literature are compared with the existing standard limitations used by the Royal Australian Air Force. Another complementary report [1] documents the general principles which should be considered when interpreting published data on POD extracted from the scientific literature and other technical reports. These principles were applied when reviewing the published literature for LPT for this report. Reference [1] also contains more detailed information about the use of NDT reliability and probability of detection information in general aerospace applications and current practice in the Royal Australian Air Force.

[2] "Limitation" is a term employed by the RAAF in technical publications and engineering standards relating to NDT.

2. Overview of Liquid Penetrant Testing

Liquid penetrant testing (LPT) was one of the earliest methods used for non-destructive inspection and has been a mainstay of practical NDT for many years. It is the most common NDT method (apart from visual inspection) used for the detection of surface-breaking cracks in metal components and is used extensively in the aerospace industry during production of aircraft and throughout their service life. LPT is the principal method used for NDT of turbine engine components, with over 90% of propulsion components being inspected using LPT at least once in their lifetime [5].

Penetrant inspection relies on a liquid dye penetrating into a surface-breaking defect, which then becomes visible once the excess dye has been removed. The dye may be coloured (normally red) for inspection under normal white light, or fluorescent for inspection under ultraviolet illumination (black light). Most liquid penetrant inspection processes involve the following major steps:

1. *Pre-cleaning*. The presence of contaminants on the surface of the part may prevent an effective inspection by either filling a defect, thus preventing the penetrant solution (dye) from penetrating, or contaminating the fluorescent penetrant solution and interfering with the fluorescence process. Etching of the component prior to inspection is sometimes used to remove smeared surface metal from prior machining processes, which may otherwise prevent penetration of the dye into defects.
2. *Penetrant application*. The penetrant dye is applied to the surface of the part for a dwell time which is sufficient to allow the dye to seep into any defects that are present.
3. *Removal of excess penetrant*. This usually involves rinsing the part, often using an emulsifier to assist with the removal of the excess penetrant. In some applications the penetrant may be wiped from the inspection area.
4. *Developer application*. A developer is used to draw the trapped penetrant dye out of the defect onto the component surface and also provide a background against which the defect indication will be readily visible. In fluorescent penetrant systems, the developer can also promote more efficient fluorescence by controlling the penetrant film thickness.
5. *Inspection*. The inspector examines the part to look for defect indications. In the case of a fluorescent penetrant inspection, the inspection is conducted in a dark room under ultraviolet light.
6. *Post-cleaning*. Residual penetrant and developer are removed.

Some penetrant inspection systems involve additional steps such as drying of the penetrant. A fluorescent penetrant system may consist of a set of specific penetrant, emulsifier and developer chemicals, which are designed to be used with a particular frequency light source and with carefully specified process steps. The timing for each step is important and is dependent on the specific penetrant materials used. For a more comprehensive overview of the practice and science of LPT refer to references [6, 7, 8].

Table 1 gives the standard classification system for penetrant processes and materials as promulgated by the American Society for Testing and Materials (ASTM), Standards Association of Australia and the RAAF [9, 10, 11].

Visible penetrants (types 2 and 3) are rarely used in modern aerospace applications and thus within the aircraft maintenance environment liquid penetrant is often considered

Table 1 Classification of penetrant materials and processes [9, 10, 11]

Penetrant	
Penetrant Type	
Type 1	Fluorescent dye
Type 2	Colour contrast (red dye)
Type 3	Visible and fluorescent dye (dual mode)
Removal method	
Method A	Water washable
Method B	Post-emulsifiable, lipophilic
Method C	Solvent removable
Method D	Post-emulsifiable, hydrophilic
Sensitivity	
Level 1	Low
Level 2	Medium
Level 3	High
Level 4	Ultra-high
Developers	
Form *a*	Dry powder
Form *b*	Water soluble
Form *c*	Water suspendable
Form *d*	Solvent-based or Nonaqueous for Type 1 fluorescent penetrant
Form *e*	Specific application or Nonaqueous for Type 2 visible dye
Solvent removers	
Class 1	Halogenated
Class 2	Nonhalogenated
Class 3	Specific application

synonymous with fluorescent penetrant testing (FPT). Many penetrant inspections are performed using a processing line for batch processing of a large number of parts. Processing lines often use a Method D post-emulsifiable hydrophilic penetrant system.

Reference [12] gives a detailed description of modern automated penetrant processing lines. Solvent-removable penetrant systems (Method C) are usually supplied in portable set of aerosol cans which can be taken to an aircraft for a field inspection of a small localised area on a component.

This literature review targets the LPT methods most commonly used on ADF aircraft, comprising [13]:

- fluorescent dye (type 1), post-emulsifiable hydrophilic method (method D), with a high or ultra-high sensitivity (level 3 or 4) and a dry powder (Form *a*) developer.
- fluorescent dye (type 1), solvent-removable method (method C), with a high or ultra-high sensitivity (level 3 or 4) and a nonaqueous (Form *d*) developer.

The existing RAAF standard limitations for these two methods are 3 mm surface-breaking length for post-emulsifiable penetrant inspection and 2 mm for solvent removable penetrant inspection.

3. Literature Review

3.1 Early Probability of Detection Trials for Liquid Penetrant Testing

The reliability of non-destructive inspection has been a topic of concern for at least forty years and liquid penetrant testing was examined in the very earliest studies. In 1968, Packman et al. [14] reported on an in-depth investigation into the reliability of the four NDT methods in regular use for crack detection at that time: radiography, dye penetrant, magnetic particle and ultrasonics. They assigned a "reliability index" to each method which was a product of the sensitivity, accuracy to determine crack size, and accuracy to determine crack location. The experimental results did not show an increasing trend in reliability with crack size. The term "minimum detectable flaw" first appears in literature of this era and "probability of detection" (POD) a few years later [15].

NDT reliability was first quantified using POD as a function of defect size in a Martin Marietta Aerospace project funded by NASA to establish design allowable flaw sizes for the NASA Space Shuttle program in the early 1970s [16]. This project investigated the performance of a wide range of inspection techniques including ultrasonics, fluorescent penetrant, radiography, acoustic emission and eddy current. The techniques were optimised for the detection of tightly closed cracks in 2219 T-87 aluminium alloy. The design-allowable flaw sizes adopted by NASA were based on a demonstration of 95% POD with 95% statistical confidence. The raw hit/miss data obtained during this study have subsequently been reanalysed by Rummel and Matzkanin using modern analysis techniques for inclusion in the Non-Destructive Evaluation (NDE) Capabilities Data Book [17], which will be discussed in Section 3.2.

In the 1970s, the Lockheed Georgia Company, on behalf of the US Air Force Logistics Command, conducted a very significant program to determine the reliability of NDT in the USAF [18]. This study involved 21 different Air Force bases and 300 US Air Force NDT technicians and was probably the largest NDT reliability assessment exercise in history, which became known colloquially as the "Have Cracks Will Travel" study. This study was specifically designed to support the NDT reliability assumptions that were specified in MIL-A-83444 "Airplane Damage Tolerance Design Requirements" [19], which assumed that a 0.5 inch in-service flaw would have 90% POD demonstrable with 95% statistical confidence. Unfortunately, the program failed to demonstrate 90% POD with 95% confidence for *any* defect size using typical inspection techniques applied by the average technician. "*With one exception, the NDI techniques employed in the program demonstrated considerable difficulty achieving a 50 percent probability of detection with 95 percent confidence for ½ inch crack sizes*" [18].

The "Have Cracks Will Travel" study used a selection of ex-service aircraft components and specimens simulating aircraft components which contained laboratory generated fatigue cracks. Teardown of the components after the POD trial provided accurate information about the true defect population. There are no details provided in reference [18] regarding the penetrant inspections applied, but it may reasonably be assumed that the inspections were performed using the standard processes in routine use at each USAF base at the time. The penetrant inspection results gave a best estimate 90% POD for defects of size a_{90} = 8.9 mm (0.35 inch) but the lower 95% confidence limit did not

Table 2 Average Flaw Detection Performance by Penetrant Type and Sensitivity Level [21]

Penetrant Type and Sensitivity Level	% False Calls	Point estimate of POD for crack length range 0.101 – 0.150 inch	
		Estimated POD (%)	Lower 95% confidence limit (%)
Type 2	0.7	89.4	81.5
Type 1, Level 2	3.3	98.3	92.5
Type 1, Level 3	9.5	94.1	88.6
Type 1, Level 4	19.8	87.1	78.0

reach 90% for any defect size. In fact, the lower 95% confidence limit only approached 60% POD for a 20 mm defect.

Despite these poor results, no changes were made to the detectable defect sizes assumed for the purpose of damage tolerance analysis, and the current standard, JSSG-2006, recommends exactly the same values as the 1974 standard, MIL-A-83444. However, the results from the "Have Cracks" study triggered corrective measures in USAF training programs to improve the POD through better technician performance, better equipment and better procedures [20].

An extensive NDT reliability assessment program was undertaken by Martin Marietta during the 1980's for detection of fatigue cracks in the high-strength alloys, Inconel® 718 and Haynes® 188 [21]. These trials included reliability assessments for liquid penetrant, eddy current and ultrasonic inspection methods. Reference [21] is significant because it includes direct comparison of the different forms of liquid penetrant testing. However, the statistical analysis of the hit/miss POD data in this report is simplistic. A point estimate of POD was found by dividing the data into crack length intervals. Confidence limits were applied to that estimate using binomial statistics, and POD curves were generated as a moving average with defect size.

The Martin Marietta study included systematic comparison of the different penetrant types, sensitivity levels, method and form of developer. Table 2 gives an extract of data from reference [21] examining the performance of different penetrant types and sensitivity levels, which show that the false call rate increases systematically with increased sensitivity level but the estimated POD does not! Indeed, for Type 1 penetrants, the reported POD for the crack length range 0.101 – 0.150 inch actually *decreased* with increasing penetrant sensitivity level.

The data from this 1988 Martin Marietta study were reanalysed by Rummel and Matzkanin for inclusion in the NDE Capabilities Data Book [17]. However, whilst the data were reprocessed using more sophisticated statistical analysis, the data were subdivided to give separate POD curves for the different methods, technicians and institutions, resulting in 29 different POD curves presented for the water-washable variant alone. This makes it difficult to draw any conclusions about the average performance of LPT. Data for other variants of LPT examined by the 1998 Martin Marietta study from reference [21] are not included in reference [17].

3.2 NDE Capabilities Data Book

The NDE Capabilities Data Book [17] was compiled by Rummel and Matzkanin with an intention to condense the available probability of detection reference data into a single source. For the 1974 and 1988 Martin Marietta studies [16, 21] discussed in Section 3.1, Rummel and Matzkanin reanalysed the raw inspection results using somewhat newer statistical analysis techniques, fitting a log-logistic model to the hit/miss data using a least-squares method. However, they do not report any confidence limits for the fitted POD curves.

For each study included in the NDE Capabilities Data Book, the data sets are subdivided into the smallest possible groupings of inspection data, regardless of the groupings used by the original researchers. This means that, rather than presenting a small number of POD curves supported by large data sets, a large number of individual POD curves are presented, each based on a relatively small data set. For example, ninety-two different POD curves are presented for fluorescent penetrant inspection, where the data originate from just six separate POD studies. Figure 3 shows a frequency plot of the a_{90} defect sizes at which the estimated POD reaches 90% for all of the fluorescent penetrant data, excluding six cases for which the log-logistic functional form assumed for the POD curve did not fit the data. The smallest reported a_{90} value was just 0.3 mm, and the largest was 17.8 mm. However, for fourteen of the 92 POD curves presented in the NDE Capabilities Data Book, the POD did not reach 90% even at the maximum crack length plotted on the graphs of 19mm (0.75 inch). In these fourteen cases, the maximum estimated POD was only of the order of 60% for a 19 mm crack, so for these fourteen cases the actual a_{90} values greatly exceeded 19 mm. Thus, for these fourteen POD curves, the only available information is that a_{90} >> 19 mm; these values are represented on Figure 3 by the bar with red hatching plotted at 19 mm defect size, representing the lower bound on the actual a_{90} values which are considerably larger, but whose precise numerical values are unknown.

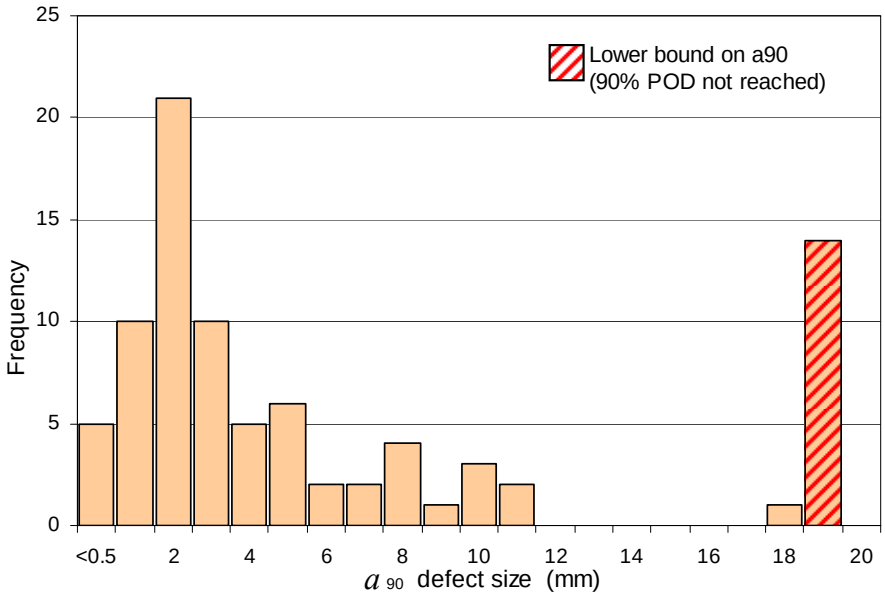

Figure 3 Frequency distribution of estimated a_{90} defect sizes for fluorescent penetrant inspection data presented in the NDE Capabilities Data Book [17]. (Excludes six data sets for which the assumed log-logistic model did not fit the data.)

3.3 Research led by Canada's NRC Institute for Aerospace Research from 1990 to 2000

The National Research Council of Canada Institute for Aerospace Research (NRC IAR) undertook a series of studies designed to establish the reliability and sensitivity of NDT used by engine maintenance organisations in Canada to inspect gas turbine engine compressor discs. The results of an initial trial (which included LPT and other methods) were summarised in a later report as:

> *"The results were surprisingly poor, and showed that the methods then employed were not able to support damage tolerance based life management of these components in a manner consistent with the stringent requirements laid down by MIL-STD-1783. Interestingly, this was … the first rigorous demonstrator project performed using real service exposed components containing service induced cracks. In contrast, most of the demonstrator projects performed to establish compliance with MIL-STD-1783 have been performed using laboratory prepared coupons containing laboratory induced cracks." [22]*

NRC IAR then coordinated a POD demonstration involving six laboratories in four NATO countries [22, 23]. A number of NDT methods were applied to the detection of fatigue cracks which had initiated and grown during service at boltholes in ex-service compressor disks and spacers from J85-CAN40 engines which had been retired due to reaching their maximum allowable service lives. Five of these laboratories applied liquid penetrant inspection to the disks and all of these used a Type 1, fluorescent penetrant of either high or ultra-high sensitivity (level 3 or 4). Table 3 shows the results for each of the laboratories. Each laboratory inspected a different set of components and the sample size, n, is the number of defects present in the disks actually inspected by each laboratory. With the exception of one 4.6 mm crack which was missed by laboratory I, all cracks larger than 2.6 mm were detected by the penetrant inspections. Four of the five laboratories (I, II, III & IV) gave similar results. However, the defect distribution in the samples inspected by laboratory VI was very different to the other laboratories as it happened to contain only small (< 2 mm) cracks, which may influence the small a_{90} demonstrated by laboratory VI. There is no apparent trend in either a_{90} or false call rate with the sensitivity of the penetrant applied.

Confidence limits on POD were computed by NCR IAR using a statistical analysis method described in reference [24], but the method used has subsequently been shown to be incorrect for analysis of hit/miss probability of detection data, with a significant

Table 3 Outcomes from NATO Round Robin Test (1995) [22 , 23]

Laboratory	Method	n	a_{90}	False call rate
I	Method B, Level 4	285	3.3 mm	0.87%
II	Method D, Level 4	404	2.5 mm	0.85%
III	Method D, Level 3	207	2.7 mm	2.6%
IV	Method D, Level 3	285	3.2 mm	0.87%
VI[3]	No information	134	1.8 mm	0.0%

[3] Defect distribution for organization VI was very different to other laboratories as it happened to contain only small (< 2 mm) cracks.

risk that the computed confidence limits will be non-conservative and invalid for the intended 95% confidence level [25, 26]. Because of this deficiency with the analysis, the 95% confidence limit defect sizes, such as $a_{90/95}$ values reported in references [22, 23, 27–29], are not considered further in this literature review. Data from this NATO round robin trial [22, 23] were incorporated into the NDE Capabilities Data Book [17], but in doing so, the original data was re-processed and analysed in groupings of laboratory and component type (either spacers or compressor disks) which differed from the groupings used in the original reports.

A subsequent study by NRC IAR examined the performance of Canadian aeroengine manufacturers, users and maintenance contractors [27, 29]. Three participating organisations (including two aeroengine maintenance contractors) applied penetrant inspections to a similar set of compressor disks as used in the above NATO study and the results are summarised in Table 4. [4] One organisation (A) had an estimated a_{90} value which was comparable to the outcomes from the NATO study, but the other two were significantly higher. Details are not provided for the different laboratories participating in the NATO study or the organisations participating in this study using Canadian manufacturers and maintainers. However, it could be inferred from the descriptions that for the NATO study the inspections may have been performed by specialist technicians in a research laboratory environment, whereas the maintenance contractors' inspections were more likely to be routine NDT performed in a high throughput production environment. Reference [27] proposes that the poor performance may be related to crack closure and surface condition.

An additional and unusual pre-processing was applied to the POD data for organisations A and D. The authors of reference [27] believed that when cracks with less than 0.3 mm surface length were detected, then the indication was unlikely to be due to the crack but were probably false calls which happened to occur at the same location as a very small crack. Consequently, hits for cracks less than 0.3 mm in surface length were excluded from the POD analysis for organisations A and D only. Excluding hits obtained on small defect sizes will tend to result in a more steeply rising POD curve and will often reduce the estimated a_{90} value. Thus, the a_{90} value reported for organisations A and D could be biased on the low side by the additional pre-processing applied to the data. Data from organisation C showed no benefit from excluding small hits from the analysis and the pre-processing to exclude small hits was not applied to the organisation C data.

A conference paper published in 1998 [28] reported only the results for organisation A, quoting an $a_{90/95}$ value of 2.59 mm, which provides a good example of selective

Table 4 Outcomes from Canadian Round Robin Test (1996) [27, 29]

Organisation	Method	n	a_{90}	False call rate
A	Method D, Level 3	246	2.3 mm	4.0%
C	Method not specified, Level 4	320	3.9 mm	2%
D	Method not specified, Level 4	318	5.7 mm	0%

reporting of results in conference papers, as discussed in reference [1].

Further examination of the data from these trials conducted by NRC IAR is presented Section 4.2, which looks in more detail at variation between organisations and presents an overall POD for data pooled across the different organisations.

3.4 Research by the USA Center for Aviation Systems Reliability from 1996 to 2008

In 2001, the Federal Aviation Administration (FAA) Center for Aviation Systems Reliability (CASR) funded a major program to perform an engineering assessment of fluorescent penetrant inspection [30]. This was primarily in response to recommendations from the National Transportation Safety Board following an uncontained engine failure during the take-off roll of Delta Air Lines flight 1288 in Pensacola, Florida on July 6, 1996. The accident investigation found that a contributing cause of the accident was the failure of the fluorescent penetrant inspection process to detect a fatigue crack which should have been readily detected [31]. The investigation recommendations included:

- "*Establish and require adherence to a uniform set of standards for materials and procedures used in the cleaning, drying, processing and handling of parts in the fluorescent penetrant inspection process. In establishing those standards, the FAA should do the following:*
    - *Review the efficacy of drying procedures for aqueously cleaned rotating engine parts being prepared for fluorescent penetrant inspections*
    - *Determine whether flash drying alone is a sufficiently reliable method*
    - *Address the need to ensure the fullest possible coverage of dry developer powder, particularly along hole walls*
    - *Address the need for a formal system to track and control development times*
    - *Address the need for fixtures that minimize manual handling of the part without visually masking large surfaces of the part.*
    - *Require the development of methods for inspectors to note on the part or otherwise document during a non-destructive inspection the portions of a critical rotating part that have already been inspected and received diagnostic followup to ensure the complete inspection of the part.*
- *Conduct research to determine the optimum amount of time an inspector can perform non-destructive testing inspections before human performance decrements can be expected.*
- *In conjunction with industry and human factors experts, develop test methods that can evaluate inspector skill in visual search and detection across a representative range of test pieces, and ensure proficiency examinations incorporate these methods and are administered during initial and recurrent training for inspectors working on critical rotating parts.*
- *Require that all heavy rotating titanium engine components (including the JT8D-200 series fan hubs) receive appropriate non-destructive testing inspections (multiple inspections, if needed) based on probability of detection data at intervals in the component's service life, such that if a crack exists, but is not detected during the first inspection, it will receive a second inspection before it can propagate to failure; assuming*

[4] The third Canadian facility providing inspection data for [27] was NRC IAR.

> *that a crack may begin to propagate immediately after being put into service, as it did in the July 6, 1996, accident at Pensacola, Florida, and in the July 19, 1989, United Airlines accident at Sioux City, Iowa."* [31]

Research to address these recommendations commenced with a review of literature on factors affecting the sensitivity of liquid penetrant inspections [8]. This comprehensive review was undertaken with the intent of identifying and organising the body of work that had led to existing liquid penetrant inspection practices. It addressed properties of penetrant materials, cleaning methods, effects of metal smear and benefits of etching, quality control, and human factors in the inspection. Although reference [8] does not explicitly address probability of detection or review outcomes of previous POD trials, it provides a valuable summary of many important findings with regard to factors that influence the performance of LPT.

One observation that has particular relevance to POD assessment of LPT is the finding that the effectiveness of a penetrant was significantly reduced if the part had been previously inspected with a different penetrant, even though the proper pre- and post-cleaning (degreasing) operations had been performed. This means that the results of round-robin POD trials, where the same set of specimens are inspected in turn by a variety of inspection laboratories, could be adversely impacted by penetrant remaining in the defect from previous inspections.

The CASR program on engineering assessment of fluorescent penetrant testing identified twelve separate engineering studies (ES) to be conducted [32].

ES 1. Developer studies [33, 34, 35]
ES 2. Cleaning studies for Ti, Ni and Al [36, 37]
ES 3. Stress studies
ES 4. Assessment tool for dryness and cleanliness
ES 5. Effect of surface treatments on detectability
ES 6. Light level studies
ES 7. Detectability studies
ES 8. Study of prewash and emulsification parameters
ES 9. Evaluation of drying temperatures
ES 10. Part geometry effects
ES 11. Penetrant application studies
ES 12. Relationship of part thickness to drying method

A major investigation into cleaning and drying processes in preparation for fluorescent penetrant inspections, resulting in 18 separate observations and recommendations, was reported in 2004 [37]. The study considered the effects on indications arising from both chemical cleaning methods and also mechanical or blasting cleaning methods. It found that effective cleaning methods exist for removing oil contamination for both Ti and Ni alloys, but that the use of some alkaline cleaners lead to reduction in FPT response. The report recommends that blasting with wet glass bead, plastic media at 80 psi, or larger (240 and 320) grit Al_2O_3 be discontinued as these cleaning methods led to surface damage and loss of penetrant indications. The study found no significant differences between flash drying at 150°F and oven drying at 225°F.

The developer studies included POD trials to compare different methods of applying Form *a* dry powder developer in a type 1, method D hydrophilic post-emulsifiable

process (see Table 1) [33, 34, 35]. A "dip and drag" method of applying the developer was compared to the cloud chamber often used in penetrant processing lines. The probability of detection was much better for the dip and drag method of applying developer and gave a_{90} = 2.2 mm (0.085 inch) and $a_{90/95}$ = 3.6 mm (0.14 inch). The POD for the cloud chamber was poor and did not reach 90% within the range of defect sizes considered (a_{90} > 5.1 mm), whilst the confidence limit value, $a_{90/95}$, would be considerably larger again. There were 16 specimens used in this trial, each 102 mm x 406 mm. The total number of cracks in the specimens was not specified in the reports, but was "*sufficient to comply with direction in the military handbook governing empirical POD studies*" [33]. References [33, 34, 35] also examine the probability of detection as a function of the brightness of the indications. The only literature on these developer studies that were found during this review were the three conference papers [33, 34, 35].

No interim reports or technical papers on the other ten planned engineering studies for the CASR program (items ES 3 to ES 12 above) have been found. The final report on the CASR engineering assessment of fluorescent penetrant inspection program was scheduled to have been published in early 2009 [35].

3.5 Studies Comparing POD for Liquid Penetrant Testing and Sonic Infrared Imaging

Sonic infrared (IR), or sonic thermography, is an advanced NDT method purported to have many advantages over LPT for inspection of engine components such as turbine blades. The potential benefits include reduced inspection time and minimal part preparation. A number of studies have been conducted to compare the probability of detection of fatigue cracks in turbine blades using sonic IR and conventional LPT.

DiMambro et al. conducted a study using titanium and Inconel® specimens which contained fatigue cracks generated by a three-point bending process [38]. The inspections were conducted using a type 1, method D penetrant process with a level 4 sensitivity. Seven ASNT qualified technicians participated, of whom two were level 1 inspectors, one was a level 2 inspector and four were level 3 inspectors. The results from each of the inspectors were analysed separately. For all of the inspectors, the probability of detection of cracks using penetrant showed little dependence on defect size. As such, even the best of the inspectors failed to demonstrate 90% probability of detection within the range of the defect sizes addressed in the study, which was up to 3.6 mm (0.14 inch) surface length. In addition, the false call rates were extremely high. One technician made 235 calls on 144 specimens, of which just 22 were calls on genuine fatigue cracks but 15 cracks were missed by this technician. The lowest false call rate was for a technician who made 34 calls on 118 specimens, of which 24 were genuine fatigue cracks but 13 cracks were missed. The best performing technician participating in this trial *almost* reached 90% POD at 3.6 mm. However, the worst performing technician achieved a POD of only 60% POD at 3.6 mm. Overall, the study concluded that sonic IR had a better POD than LPT for defects greater than 1 mm (0.040 inch) in length. Defects less than 1 mm in length were more likely to be detected by LPT than sonic IR, but at the expense of a high false call rate [38].

An extensive investigation by Mayton into the performance of sonic IR also compares sonic IR results with those obtained by penetrant inspections [39]. This study found that cracks with less than 0.7 mm surface length were not detected by fluorescent penetrant.

Neither method was characterised using traditional probability of detection in this study.

3.6 Other Recent Research in POD for Liquid Penetrant Testing

Brausch and Tracy conducted a study on the effect of compressive stress on fluorescent penetrant indications of fatigue cracks in titanium [40]. This study found that compressive stress reduces the area and fluorescent intensity of crack indications and made the following recommendation:

> "*Fluorescent penetrant inspection should not be utilized for inspection of titanium (Ti-6Al-4V) components in the following situations:*
> *a) compressive surface stresses are expected to be above 40 ksi and*
> *b) detection of flaws with lengths of 0.060 inch* [1.5 mm] *or less is required.*" [40]

All the cracks used in the investigation had surface lengths less than 1.5 mm (0.060 inch). The report draws no conclusions regarding the detectability of cracks longer than 1.5 mm, and in fact recommends: "*Fluorescent penetrant inspection should not replace eddy current inspection without the benefit of well-designed POD studies on components containing residual stresses typical of in-service components*" [40].

Research into the effect of operator fatigue on fluorescent penetrant inspections has been sponsored by the FAA in response to the findings of the investigations into accidents at Sioux City and Pensacola. (Those recommendations were discussed in Section 3.4.) Drury et al report on a factorial experiment aimed at identifying the most significant interactions among key human factors variables such as time on the task, number of breaks and day or night shifts [41]. This study characterised performance using a probability of detection of defects but did not consider the effect of defect size, thus POD was calculated as the total number of defects detected (hits) divided by the total number of defects inspected (binomial proportion).

Aerospace and rocket engine manufacturer Pratt & Whitney requires that inspectors performing penetrant inspections on fracture critical components periodically undertake a personal POD assessment to ensure that they are achieving adequate reliability. Each inspector is assessed for each inspection system on which they work. Lively and Aljundi compiled data obtained during these assessments in the period 1998-2002 and provided the set of $a_{90/95}$ values demonstrated by each inspector each year [42]. For each $a_{90/95}$ value given, the penetrant product is specified but other details of the inspection process are not provided. The paper indicates that a range of processes have been assessed:

> *"POD data has been collected on a number of different inspection lines. These lines range from large capacity commercially manufactured lines to small in house manufactured lines. Penetrant application has been accomplished by a number of methods including electrostatically spraying, immersion, painting and spraying. Washing has been accomplished by manual spraying, automated mechanical spraying, manual immersion and automated mechanical immersion. Emulsifier has been applied by manual immersion, automated mechanical immersion and by manual spraying. Drying has been accomplished using manufactured* [fluorescent penetrant inspection] *FPI ovens, heat treat ovens and bread warming ovens. Developer has been applied by manual dusting, cloud chambers, developer wands and non-aqueous spray."* [42]

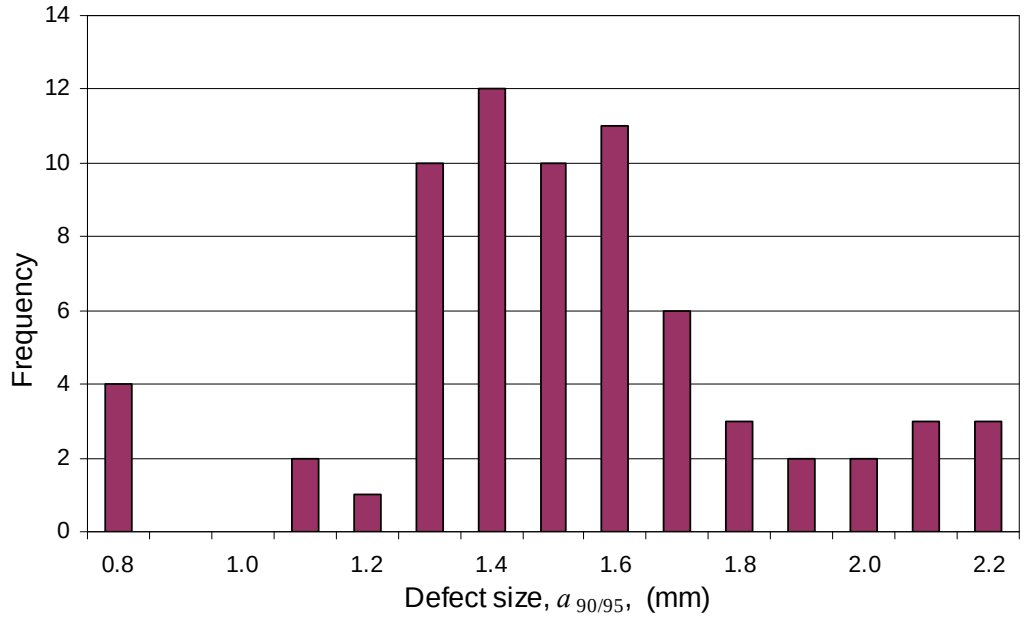

Figure 4 Frequency distribution of $a_{90/95}$ defect sizes demonstrated during FPI inspector assessment POD trials conducted by Pratt & Whitney over the period 1998 to 2002. (Data taken from tables presented in reference [42].)

Figure 4 shows a frequency histogram of the $a_{90/95}$ values obtained during the Pratt & Whitney POD assessments, as presented in the tables in reference [42]. These data are very valuable because they cover a wide range of technicians and inspection processes over an extended period. However, they may indicate the best performance of the inspection system, rather than average performance, because technicians knew that they were individually being assessed during this exercise and they were observed throughout the inspection process. In addition, the specimens used were flat plates, rather than the often complex geometry of more typical aerospace components.

Reference [42] includes some insightful discussion regarding process errors commonly encountered during these trials. Observations included "*Inspectors not accustomed to seeing very small flaws have concluded that the panels do not contain any flaws even though there are fluorescent indications visible. Some inspectors have missed large flaws because they conclude large fluorescent indications can not possibly be flaws. Some inspectors have concentrated so much on finding small flaws that they have overlooked the large flaws.*"

One concerning comment made by the authors is that "*POD panel sets must be calibrated to produce the same results for the same process. By selectively removing flaws from the analysis, panel sets can be calibrated to provide the same results.*" [42]. This seems to imply a lack of understanding of the random nature of detection and suggests that flaws which are on the edge of detectability may have been removed from the analysis when, in fact, they could provide the best information about the performance of the system.

4. Analysis of Detectable Defect Sizes from Published Literature

This review has identified twelve separate major studies on the reliability of liquid penetrant testing conducted between 1968 and 2009. Reports on these activities contained over 150 different published detectable defect sizes, including many demonstrated a_{90} or $a_{90/95}$ values.

Table 5 summarises the outcomes of POD trials discussed in Section 3 in terms of a minimum reliably detectable defect size, a_{NDI}, that might be justified based on each study.[5] Many of these are best estimates of the a_{90} defect size at which the POD reaches 90%; others take the sampling variability into account with a 95% confidence limit giving $a_{90/95}$ values.

Table 6 presents a summary of defect sizes, a_{NDI}, that various engineering standards assert can be assumed for engineering purposes to be reliably detectable by liquid penetrant inspections. These are the defect sizes that are recommended to be used in engineering analysis such as fatigue crack growth or durability and damage tolerance analyses. These defect sizes are considered to be conservative values that may be used as a default initial defect size without undertaking any special trials to validate the capabilities of the NDT techniques. Generally across the airworthiness standards, smaller defect sizes than these may be assumed if they have been demonstrated to have at least 90% probability of detection, demonstrated with 95% statistical confidence [1].

Ideally, for an equivalent defect type and inspection scenario, the a_{NDI} values given in Table 5 should be smaller than those in Table 6. Although Table 5 presents a wide range of a_{NDI} values, generally speaking the detectable defect sizes for damage tolerance analysis of aircraft structures given in JSSG-2006 (Table 6) compare reasonably to the results from POD trials in Table 5. However, the assumed detectable defect sizes for crack growth analysis in engine components given in MIL-HDBK-1783B appear to be highly optimistic compared to the results from POD trials. This literature review found no data in the public domain to support the defect sizes of 0.4 mm and 0.5 mm assumed for automated fluorescent penetrant inspections in MIL-HDBK-1783B. The procedure limitations issued with the RAAF general procedures for liquid penetrant testing have a similar size to the results from POD trials, but are more towards the optimistic end of the range of a_{NDI} values from published POD trials presented in Table 5.

The data found in the literature are generally applicable to the post-emulsifiable penetrant systems, methods B or D. There are no data specifically evaluating solvent removable (method C) penetrant inspections.

[5] Table 5 includes results only from those studies that provided a detectable defect size based on assessment of probability of detection.

Table 5 Detectable defect sizes demonstrated by probability of detection trials

a_{NDI}	Basis of a_{NDI}	Specification of Penetrant system	Source
8.9 mm	a_{90}	Application of standard processes in use across 21 USAF bases in 1970's.	Lockheed Georgia / USAF "Have Cracks Will Travel" POD trial [18]
1.8 to 3.3 mm	Range of the a_{90} values demonstrated by five different laboratories	Method B or D, level 3 or 4.	Round robin POD demonstrations for NATO laboratories coordinated by NRC IAR [22, 23]
2.3 to 5.7 mm	Range of the a_{90} values demonstrated by three different organisations	Level 3 or 4	NRC IAR round robin POD demonstration including two Canadian aeroengine maintenance contractors [27, 29]
3.6 mm	$a_{90/95}$	Type 1, method D, Level 4, Form *a*. Developer applied using a dip/drag method	CASR LPT developer study [33, 34, 35]
> 5.1 mm	a_{90} POD did not reach 90% within range of defect sizes used (maximum defect size 5.1 mm).	Type 1, method D, Level 4, Form *a*. Developer applied using a cloud chamber method	CASR LPT developer study [33, 34, 35]
> 3.6 mm	a_{90} Best performing technician almost reached 90% POD at 3.6 mm. Worst performing technician only achieved approximately 60% POD at 3.6 mm.	Type 1, method D, Level 4	POD trials conducted to compare performance of sonic IR and LPT [38]
2.3 mm	Upper limit on $a_{90/95}$ values demonstrated by individual inspectors under examination conditions	Wide variety of processes	Pratt & Whitney inspector performance validation POD trials – historical data [42]

Table 6 Defect sizes assumed for engineering purposes to be reliably detected by penetrant inspections

a_{NDI}	Nature of a_{NDI}	Specification of Penetrant system	Source
6.3 mm (0.25 inch)	Flaw size assumed to exist following an in-service inspection for aircraft structures, being either: 6.3 mm long through-thickness crack at holes with thickness $t \leq 6.3$ mm or 6.3 mm radius of a corner crack at holes with $t > 6.3$ mm	Penetrant NDT[6]	JSSG-2006 Guide to the Specification of Aircraft Structures [2] Table XXXII In-service inspection initial flaw assumptions (p450)
12.7 mm (0.5 inch)	Flaw size assumed to exist following an in-service inspection for aircraft structures, being either: 12.7 mm long through-thickness flaw in $t \leq 6.3$ mm or 12.7 mm long x 6.3 mm deep semi-circular surface flaw in $t > 6.3$ mm.	Penetrant NDT[6]	JSSG-2006 Guide to the Specification of Aircraft Structures [2] Table XXXII In-service inspection initial flaw assumptions (p450)
1.8 mm (0.07 inch)	Maximum flaw size that can exist in a part after manufacture. Applicable to aircraft engine components. This is explicitly stated to be an assumed $a_{90/95}$ for *all* manual inspection methods.	Manual fluorescent penetrant inspection	MIL-HDBK-1783B Engine Structural Integrity Program [3].

[6] The quoted a_{NDI} value is specified in JSSG-2006 as a generic value which is considered to be appropriate for in-service inspection using penetrant, magnetic particle, eddy current, or ultrasonic inspection techniques.

Table 6 Defect sizes assumed for engineering purposes to be reliably detected by penetrant inspections (continued)

a_{NDI}	Nature of a_{NDI}	Specification of Penetrant system	Source
0.5 mm (0.02 inch)	Maximum flaw size that can exist in a titanium alloy part after manufacture. Applicable to aircraft engine components. This is explicitly stated to be an assumed $a_{90/95}$ for *all* automated inspection methods.	Automated fluorescent penetrant inspection of titanium alloy parts	MIL-HDBK-1783B Engine Structural Integrity Program [3].
0.4 mm (0.014 inch)	Maximum flaw size that can exist in a nickel alloy part after manufacture. Applicable to aircraft engine components. This is explicitly stated to be an assumed $a_{90/95}$ for *all* automated inspection methods.	Automated fluorescent penetrant inspection of nickel alloy parts	MIL-HDBK-1783B Engine Structural Integrity Program [3].
2.5 mm	Minimum detected crack size given in standard	Penetrant NDT	ISO 21347 Space systems – Fracture and damage control [43]
3 mm	Procedure limitation for RAAF LPT/GEN/1 general procedure	Type 1, method A, Level 3, Form *a* or *d*	LPT/Gen/1 Water washable FPT for use on bulk processing lines [13]
3 mm	Procedure limitation for RAAF LPT/GEN/2 general procedure	Type 1, method D, Level 3 or 4, Form *a* or *d*.	LPT/Gen/2 Post-Emulsifiable Hydrophilic LPT [13]
2 mm	Procedure limitation for RAAF LPT/GEN/3 general procedure	Type 1, Method C, Level 3 or 4, Form *d*	LPT/Gen/3 Solvent removable LPT [13]

4.1 Meta-Analysis of a_{90} Values from Published Literature

Figure 5 shows the frequency of a_{90} values from the published literature discussed in Section 3 and summarised in Table 5. Figure 5 includes eight a_{90} values from the NRC IAR studies, two a_{90} values from the CASR developer study, one result from the sonic IR vs LPT comparison, and 86 results from the NDE capabilities data book.[7] It does not include the historical $a_{90/95}$ values from the Pratt & Whitney inspector performance POD validations because of the likely bias in these data, as discussed previously. For trials that failed to achieve a 90% POD, it is simply known than a_{90} is greater than some value. These lower bound values are also shown on Figure 5, denoted by the red hatching. Of all these 97 different trial results, the mean a_{90} = 5.9 mm and the median a_{90} = 3.2 mm.[8]

However, the results from the NDE Capabilities Data Book [17] should not be treated as for the other trial results. Firstly, 81 out of the 92 POD data sets for penetrant inspections are results from trials conducted in the 1970's by Martin Marietta which may have less relevance to current LPT practice. In addition, the authors of this reference reduced each data set to the smallest possible group of inspection data regardless of what the original researcher decided was an appropriate grouping. This means that the sample sizes are

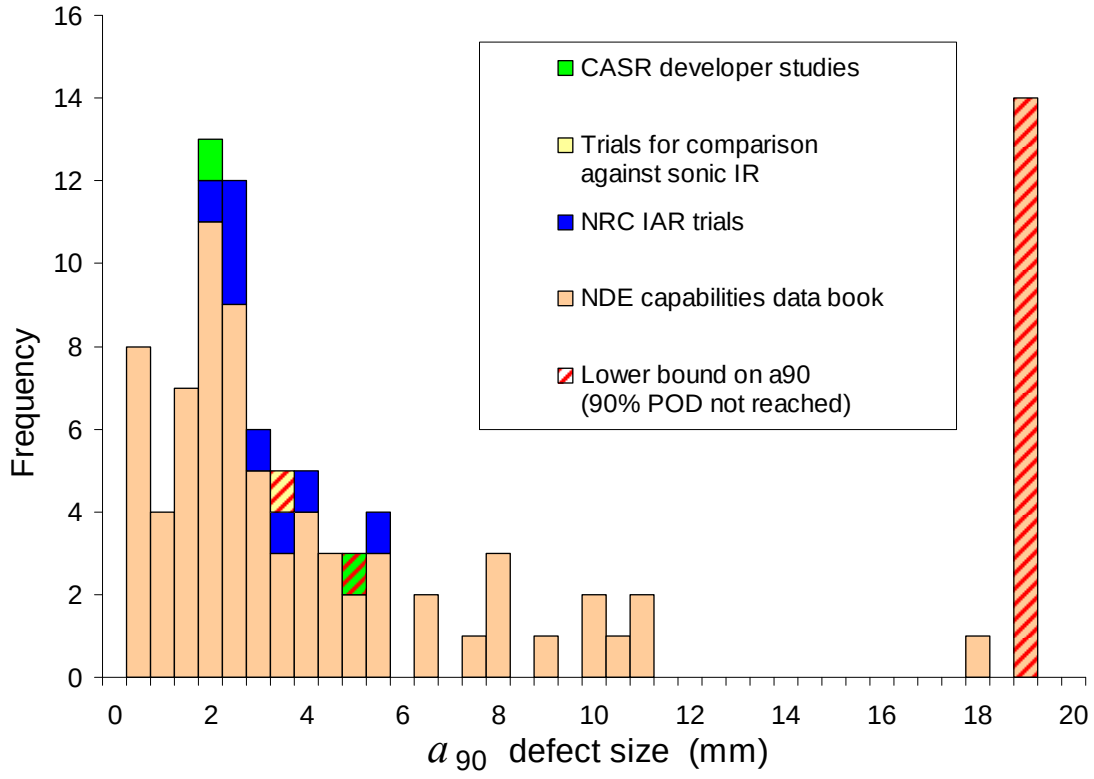

Figure 5 Frequency histogram of detectable defect sizes demonstrated by probability of detection trials for fluorescent penetrant inspection

[7] The six data sets for which a log-logistic POD curve did not fit the data are excluded from Figure 5.

much smaller for these individual data sets than for the full trial results from other research, and much greater scatter in the reported a_{90} values should be expected.

Removing the results from the NDE Capabilities Data Book from the set leaves just eleven POD trial results that are applicable to modern practice of post-emulsifiable LPT. This set includes the NRC IAR trials, the CASR developer studies and the trials conducted to compare sonic IR and LPT. A frequency histogram for these results is given in Figure 6. For these data the mean a_{90} = 3.3 mm and the median a_{90} = 3.2 mm.[9]

As discussed in Section 3, there are weaknesses with each of the individual trial results plotted in Figure 6, but taken together they provide useful information about what defect sizes might reasonably be expected to be detectable by liquid penetrant testing. If we consider that this set of eleven a_{90} values is a sample drawn from the true population of all penetrant systems consistent with modern practice, then it's possible to make inferences regarding the distribution of a_{90} for modern penetrant systems.

Maximum likelihood estimation (MLE) was used to fit a log-normal distribution to the observed data, as shown in Figure 7. The assumption of a log-normal distribution is consistent with the overall shape of the frequency distribution observed in Figures 5 and 6, and also consistent with the fact that a_{90} can only have positive values (lognormal

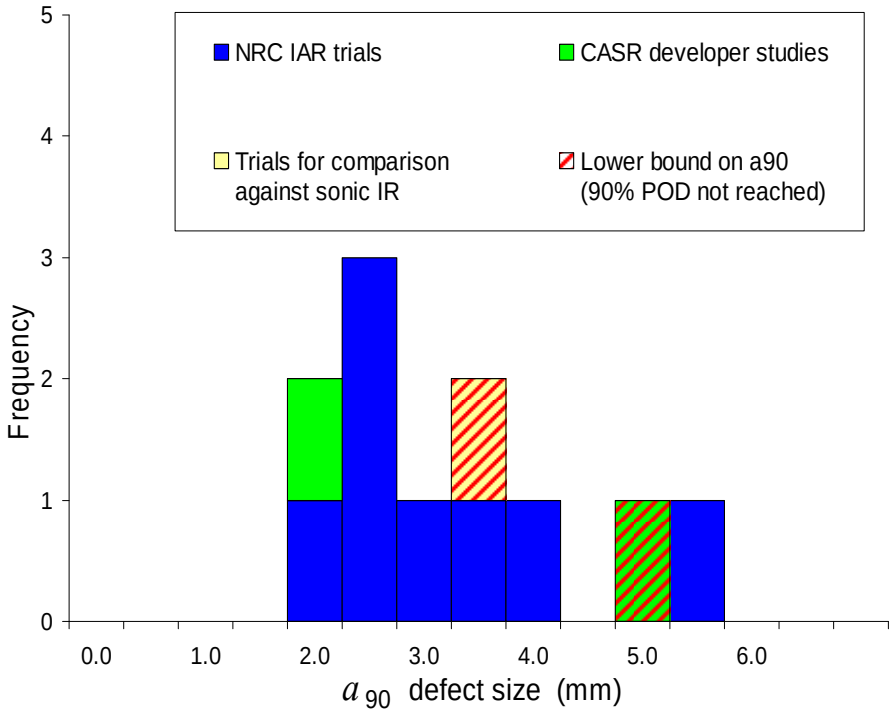

Figure 6 Frequency histogram of detectable defect sizes demonstrated by probability of detection trials for fluorescent penetrant inspection – excluding data from the NDE Capabilities Data Book[10]

[8] The lower bound values were included in the calculation of the mean and median a_{90} values.

[9] It is a coincidence that the data presented in Figures 5 and 6 both have a median of a_{90} = 3.2 mm.

[10] Red hatching (superimposed over another colour) denotes data for which the plotted value is a lower bound on the a_{90} defect size. It is known only that a_{90} exceeds this value.

ensures $a_{90} > 0$). The MLE fit allows for the data from trials where the POD did not reach 90%. (These trials provide lower bounds on a_{90} which are treated as right-censored data points in the analysis.) From this fitted log-normal distribution of a_{90} values, the *best estimate* of the median[11] a_{90} for all modern penetrant systems is $\hat{\mu}_{a90} = 3.3\,\text{mm}$. It is possible to compute a confidence interval on this median using the likelihood ratio test, with the result that there is 95% confidence that the true median μ_{a90} lies within the interval $2.5 \le \mu_{a90} \le 4.3$. Thus, an upper 95% confidence limit on the *average* (median) defect size at which the probability of detection reaches 90% is 4.3 mm (averaged across all modern penetrant systems, as represented in these eleven POD trial results).

The statistics in the previous paragraph describe the centre of the distribution of a_{90} values for different LPT systems based on the available literature. However, a statistic of more interest might be: Given the scatter in a_{90} values obtained when POD trials have been conducted, what is the largest a_{90} that might be reasonably be expected for any individual LPT system? This gives a measure of the worst case performance (largest a_{90}) that might be encountered for any individual LPT system based on the scatter in performance observed in the published data. A 95% prediction limit on future measurements of a_{90} values based on the fitted distribution is a_{90} = 6.2 mm, i.e. infer that 95% of all randomly selected LPT systems will have an a_{90} that is less than or equal to 6.2 mm. This 95% prediction limit is a best estimate of the right-hand tail of the

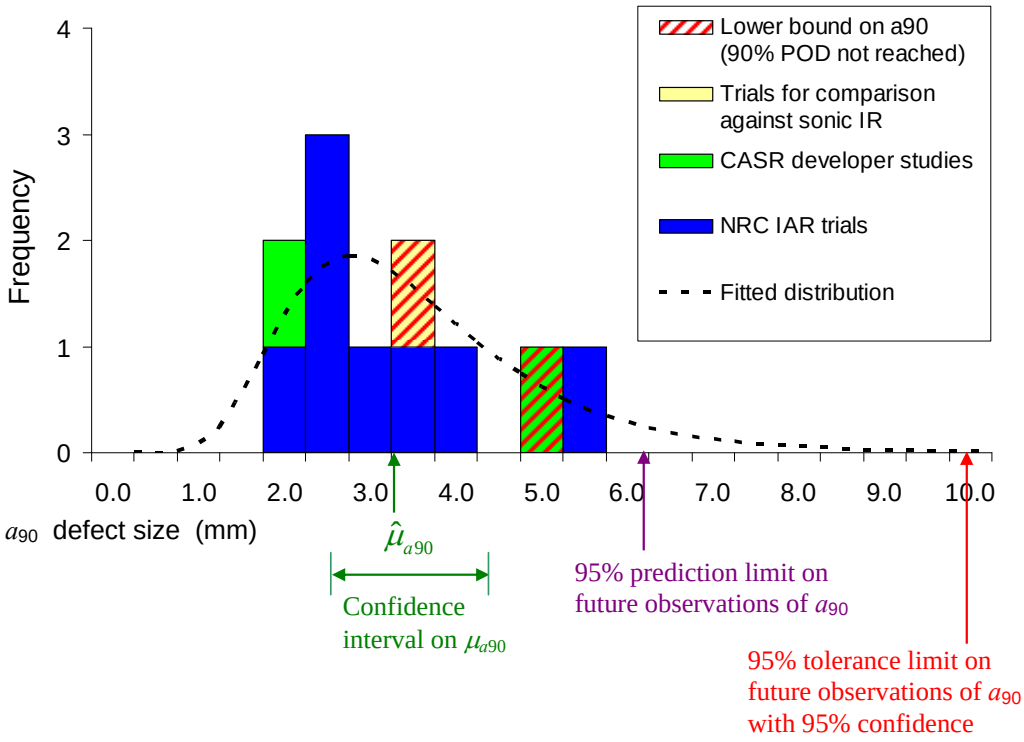

Figure 7 Frequency histogram of detectable defect sizes demonstrated by probability of detection trials for fluorescent penetrant inspection showing statistical inferences on a_{90}

[11] For the log-normal distribution, the location parameter, μ, used to describe the distribution is the median rather than the mean.

distribution, based on the best fit of the assumed log-normal distribution to the observed data (reported a_{90} values). To take into account the random effect of sampling variability on this estimate, particularly given that it is based on a small sample of just eleven different penetrant systems, confidence limits can be applied to this prediction limit. A confidence limit applied to a prediction interval is known as a statistical *tolerance interval*, which is defined as a confidence interval that contains a fixed proportion of the population at a given confidence level [44]. Applying tolerance limits to these data, we can be 95% confident that 95% of the population of modern penetrant systems will have an a_{90} value less than or equal to 10.0 mm. Figure 7 displays these four different statistical inferences on a_{90} graphically.

Of the eleven POD trial results that are applicable to modern practice of post-emulsifiable LPT, those conducted by NRC-IAR are considered to be the most relevant. The POD trials administrated and reported on by NRC-IAR [22, 27] used retired engine components and collected data across a variety of organisations, including engine manufacturers and maintainers. Figure 8 shows a frequency histogram of the eight a_{90} defect sizes from the POD trials conducted by NRC-IAR, along with a fitted log-normal distribution and statistical inferences on a_{90}. Table 7 shows the statistical inferences both for the full set of eleven a_{90} defect sizes and for just the set of eight a_{90} values from the NRC-IAR trials. Prediction and tolerance limits are computed at both 90% and 95% levels.

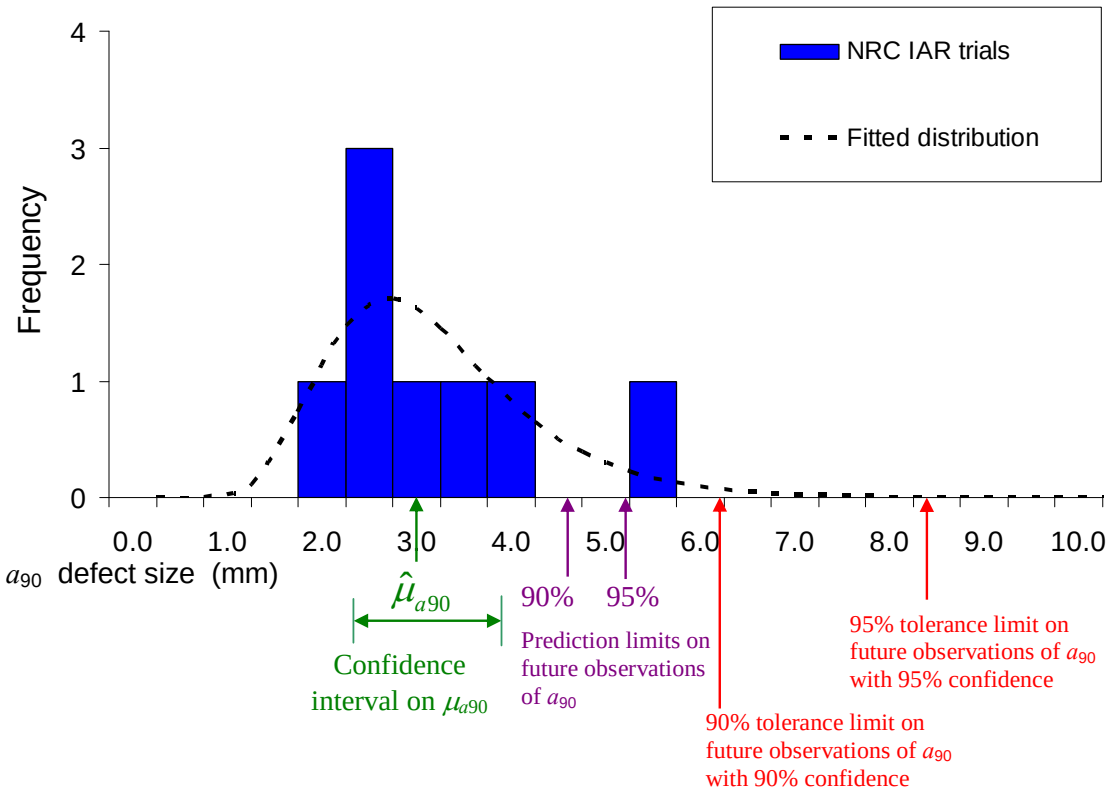

Figure 8 Frequency histogram of eight a_{90} detectable defect sizes from NRC IAR data [22, 27] showing associated statistical inferences on a_{90} given in Table 7.

Table 7 Statistical inferences on a_{90} *(mm) for NRC IAR POD trial results as plotted in Figure 8, compared to eleven trial results plotted in Figure 7*

	Using eight trial results from NRC IAR trials [22, 27]	Using eleven trial results from Figure 7 including NRC IAR trials, CASR developer studies and trials for comparison against sonic IR.
$\hat{\mu}_{a90}$	3.0	3.3
Upper confidence limit on $\hat{\mu}_{a90}$	3.9	4.3
90% prediction limit	4.6	5.4
95% prediction limit	5.2	6.2
90% tolerance limit with 90% confidence	6.1	7.1
95% tolerance limit with 95% confidence	8.4	10.0

None of the statistics in Table 7 are strictly equivalent to the $a_{90/95}$ value that would be obtained from a conventional POD trial. The principles underpinning each of the four inference statistics are described in Appendix A, where they are separated into:

- statistics relating to the average performance across all implementations of post-emulsifiable LPT; and
- statistics that consider that LPT performance will vary between facilities and anticipate the worst performance of LPT likely to be encountered.

4.2 Analysis of Raw Hit/Miss POD Data for NRC IAR Studies

The NDE Capabilities Data Book [17] discussed in Section 3.2 not only provides probability of detection curves but also has an accompanying CD which provides spreadsheets containing the raw hit/miss inspection results for each of the data sets. This enables the raw inspection results to be reanalysed to fit POD curves using modern analysis methods. As discussed previously, the NDE Capabilities Data Book divides the data sets into the smallest possible groupings of inspection data, resulting in a large number of POD curves each for a relatively small data set and with a large expected scatter in the estimated POD curves due to the small data set size. Re-analysis of the raw hit/miss inspection data can overcome this problem by appropriate pooling of the data to give POD curves which are based on larger data sets. Of the data sets available from the Data Book, only those from POD trials conducted by NRC IAR (refer to Section 3.3) are considered to be applicable to modern practice of liquid penetrant testing. This section presents the results of DSTO analysis of the raw hit/miss POD data for the NRC IAR trials to determine best estimate and lower 95% confidence limit POD curves.

The raw hit/miss POD data for LPT from the NRC IAR trials were extracted from [17] and cross-checked against the data presented in the original NRC IAR reports [22, 27]. Many discrepancies were found between the data from [17] and the data available in the original NRC IAR reports. There are also inconsistencies within the original NRC IAR reports themselves, in particular between the summary data in the body of these reports and the data provided in the appendices to [22] and [27]. Appendix B provides an overview of the inconsistencies found with these data. However, these inconsistencies represent a small proportion of the total data set available and are not expected to have a large influence on conclusions that may be drawn from analysis of the raw data.

The available hit/miss POD data for LPT from the NRC IAR trials were pooled to determine an average POD result for the entire data set. Figure 9 shows the estimated probability of detection and the lower confidence limit fitted to all available data for the eight NRC IAR trials. [12] The raw data are represented as circles showing the proportion of hits in each 0.25 mm defect size interval, with area proportional to the number of data points in the interval. The combined data set for [22, 27] contains inspection data from laboratories in different NATO countries [22] and from three Canadian facilities including two aircraft engine maintainers [27].[13] Figure 10 shows the probability of

[12] The analyses presented in Figures 9 and 10 and Table 8 fit a four-parameter POD curve to the data, rather than the simpler two-parameter curve used for previous DSTO analysis of POD data. The use of a four-parameter curve is recommended in [45] and is particularly suitable when the inspection data contains a significant rate of hits at very small defect sizes, as is the case for these data. The majority of these small hits are likely to be false calls occurring in locations which contain small cracks. If these data are used to fit a standard two-parameter POD curve, the POD curve will be distorted and the a_{90} values will be increased. The four-parameter model includes a parameter specifically to account for the false call rate, and also a parameter which describes the probability of missing very large defects (independent of defect size). Confidence limits were determined using a likelihood ratio test on the four-parameter model. DSTO analyses of the NRC IAR trial data are documented in [46].

[13] The third Canadian facility providing inspection data for [27] was NRC IAR.

Table 8 Values of a_{90} and $a_{90/95}$ for POD fitted to pooled hit/miss data from NRC IAR trials

	a_{90}	$a_{90/95}$
All available data combined for NATO laboratories (I, II, III, IV and VI from [22]) and Canadian facilities (A, C and D from [27])	2.6	3.2
Data from Canadian facilities A, C, D only [27]	3.5	5.2

detection estimated for only the Canadian facility inspection results from [27]. The a_{90} and $a_{90/95}$ statistics from the fitted POD curves are given in Table 8.

This pooled analysis of the eight raw data sets from the NRC IAR POD trials [22, 27] gives $a_{90/95}$ = 3.2 mm, representing a confidence limit on the notional average POD across the participating facilities. Analysis of the raw data for only the Canadian facilities [27] gives a substantially larger $a_{90/95}$ = 5.2 mm.

POD curves were also computed for the eight NRC IAR data sets separately [46]. Comparison of these POD curves shows a significant difference in performance between inspection facilities which cannot be explained by sampling variation alone. Given this observed variability between facilities, it would be unwise to attempt to describe the POD for all implementations of LPT via a single 'average' POD curve determined from a pooled data set. Instead, analysis of the variability observed between different LPT facilities is recommended, such as that presented in Section 4.1.

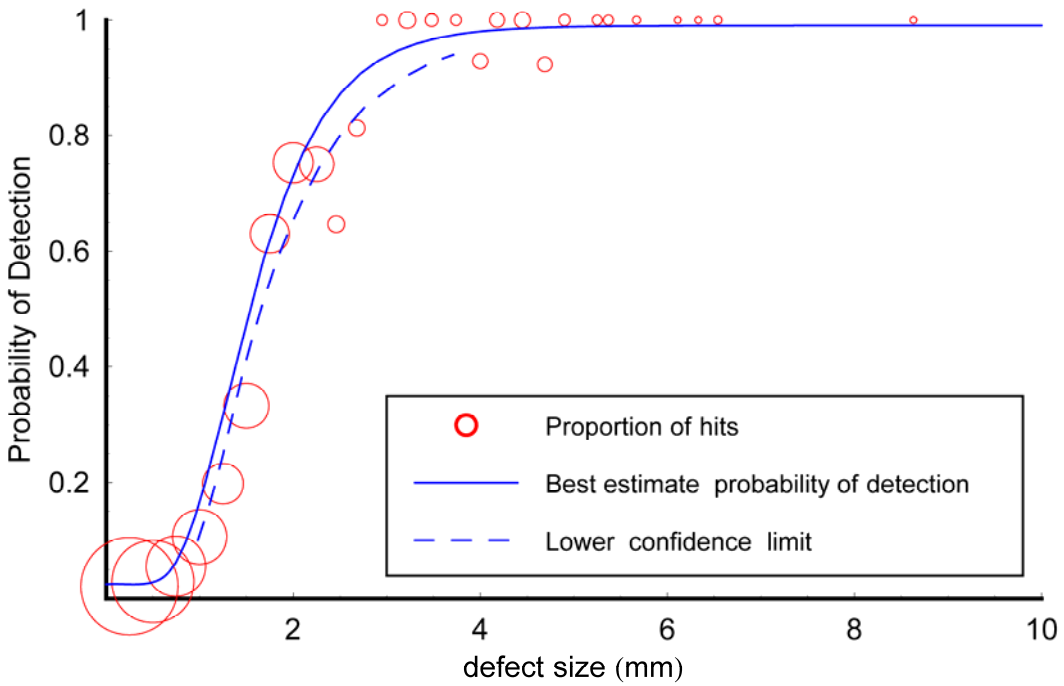

Figure 9 Estimated probability of detection for all available hit/miss data from the eight NRC IAR trials. Total number of inspection results is n = 2032.

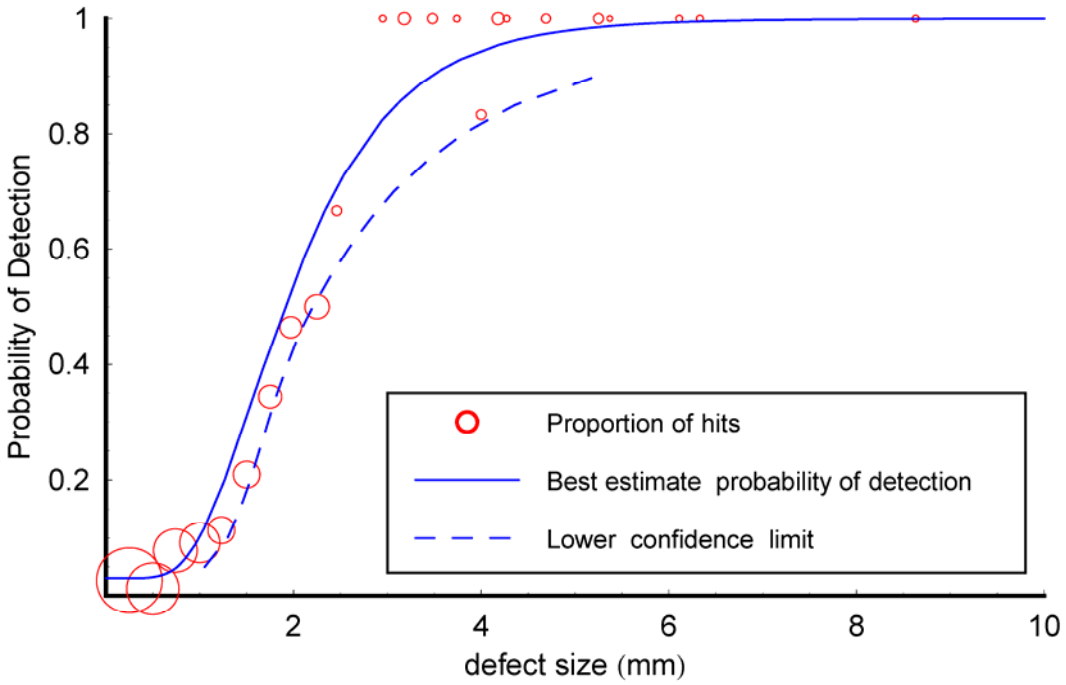

Figure 10 Estimated probability of detection for pooled raw hit/miss data for Canadian facilities only. (Organisations A, C and D from [27].) Total number of inspection results is n = 891.

4.3 Analysis Summary

The preceding sections have described the range of a_{90} values reported in the published literature, the results of a meta-analysis of those values, and the results of a POD analysis of the raw hit/miss data from the NRC IAR POD trials with the data pooled to generate a single average POD curve, for either the total data set or the data from the Canadian facilities only.

From the meta-analysis of the eight a_{90} values reported from the NRC IAR POD trials (Section 4.1), the following key statistics have been obtained:

- The best estimate of the a_{90} value which is expected to be achieved by 50% of LPT process implementations is 3.0 mm. (*Best estimate of the median,* $\hat{\mu}_{a90}$.)
- There is 95% confidence that at least 50% of LPT processes will achieve $a_{90} \leq$ 3.9 mm. (*Upper 95% confidence limit on* $\hat{\mu}_{a90}$.)
- The 90% prediction limit on a_{90} from the eight NRC IAR trial results of 4.6 mm. This means that, based on the available data, 90% of LPT process implementations are expected to achieve $a_{90} \leq$ 4.6 mm. This 90% prediction limit may provide the most appropriate basis for a_{NDI} for LPT, giving a_{NDI} = 5 mm (rounded to the nearest millimetre).

By comparison, Section 4.2 presented a pooled analysis of the eight raw data sets from the NRC IAR POD trials [22, 27] giving $a_{90/95}$ = 3.2 mm, representing a confidence limit

on the notional average POD across the participating facilities. Analysis of the raw data for only the Canadian facilities [27] gives a substantially larger $a_{90/95}$ = 5.2 mm.

There is evidence from the range of a_{90} values plotted in the figures in Section 4.1 to indicate that the variability in the performance of LPT processes between facilities is significant. This means that statistics relating to the *average* performance across the different implementations of LPT sampled in the studies are probably not an appropriate basis for an a_{NDI} for RAAF LPT procedures.

Variability between facilities is of particular importance when compared to other factors that may influence reliability such as variations in performance between different technicians, or for different materials or surface finish. Whilst these other factors can reasonably be expected to have an effect which is randomised across different inspections on each component, there is potential for the difference between facilities to adversely impact all inspections conducted at a given facility if that facility achieves a POD which is consistently below the average. This could result in a cumulative increase in the overall probability of catastrophic failure if components for a given aircraft type are typically inspected only at that particular facility.

Thus, as the evidence indicates significant differences between the performance of LPT processes at different facilities, detectable defect sizes should be based on the worst performance of LPT likely to be encountered rather than an average performance expected across different facilities.

5. Conclusions and Recommendations

A review of published literature on the reliability of liquid penetrant testing found results from twelve major studies conducted between 1968 and 2009. Reports on these activities contained over 150 different published detectable defect sizes, including many demonstrated a_{90} or $a_{90/95}$ values. This wide range of data was reviewed to identify eleven different a_{90} defect sizes that were assessed to be reasonably applicable to the modern practice of post-emulsifiable LPT. All available data relate to post-emulsifiable LPT; no data were found in the literature which specifically address reliability of solvent removable method C penetrant inspections.

The assumed detectable defect sizes for damage tolerance analysis of aircraft structures given in JSSG-2006 compare reasonably to the detectable defect sizes demonstrated by POD trials for LPT processes. However, the assumed detectable defect sizes for crack growth analysis in engine components given in MIL-HDBK-1783B are highly optimistic compared to published results from POD trials. No data were found to support the minimum detectable defect sizes of 0.4 mm and 0.5 mm recommended for automated fluorescent penetrant inspections in MIL-HDBK-1783B.

The published data indicate significant variability in the performance of liquid penetrant testing between the different implementations of the post-emulsifiable LPT process. Based on this, a similar degree of variability could be expected between the different implementations of post-emulsifiable LPT processes used on ADF aircraft.

The current a_{NDI} = 3.0 mm for LPT in accordance with RAAF general procedure LPT/GEN/2 [13] is consistent with estimates of the *average* performance of LPT demonstrated in the literature. The available data indicate that approximately 50% of implementations of post-emulsifiable LPT will have an a_{90} which exceeds 3 mm. Consequently, the available data do not support any reduction in the existing a_{NDI} for LPT and an increase in a_{NDI} may be appropriate.

Based on the data available from the published literature, the smallest a_{NDI} that might reasonably be assumed to be detected with 90% probability of detection by *most* implementations of post-emulsifiable liquid penetrant testing processes is 5 mm. This value is based on a 90% prediction limit for a_{90} derived from the eight POD trial results reported by NRC IAR in references [22, 27].

6. Acknowledgements

The authors gratefully acknowledge the assistance received from Matthew Khoo, Robert Ditchburn, Peter Virtue and Howard Morton in identifying and locating many of the references for liquid penetrant testing. Particular thanks are due to Matthew Khoo for his considerable effort to extract relevant data from the NDE Capabilities Data Book.

7. References

1. Harding, C. A. and Hugo, G. R. (2009) *Guidelines for Interpretation of Published Literature on Probability of Detection for Nondestructive Testing*. DSTO-TR-2622, Defence Science and Technology Organisation, Melbourne, Australia.

2. *Joint Service Specification Guide Aircraft Structures*. (1998) JSSG-2006, Department of Defense, USA.

3. *Engine Structural Integrity Program (ENSIP)*. (2004) MIL-HDBK-1783B, USA.

4. *Design and Technology Services Support Manual, Section 2 Chapter 9 Non Destructive Testing Standards Laboratory*. (2008) AAP 7001.068(AM1), Australian Defence Force, Australian Air Publication.

5. Brasche, L. (2002) The Use of Nondestructive Inspection in Jet Engine Applications. *Materials Evaluation* **60** (7) July, pp. 853-856.

6. Moore, P. O. and Tracy, N. A. (eds.) (1999) *Liquid Penetrant Testing*. 3rd ed. Nondestructive Testing Handbook Vol. 2, Columbus, Ohio, USA, American Society for Nondestructive Testing.

7. Borucki, J. S. and Jordan, G. *Liquid Penetrant Inspection*. ASM Handbook Volume 17 Nondestructive Evaluation and Quality Control. (2002) [Accessed 2009 23 March]; Available from: http://products.asminternational.org/hbk/index.jsp.

8. Larson, B. (2002) *Study of the Factors Affecting the Sensitivity of Liquid Penetrant Inspections: Review of Literature Published from 1970 to 1998*. DOT/FAA/AR-01/95, Federal Aviation Administration, Office of Aviation Research, Washington, DC, USA.

9. *Standard Practice for Liquid Penetrant Testing*. (2005) E1417 -05, ASTM International, USA.

10. *Non-destructive testing - Penetrant testing of products and components*. (1997) AS 2062 - 1997, Australian Standard, Standards Association of Australia.

11. *Non Destructive Inspection Methods General Data*. (1992) AAP 7002.008, Defence Instruction (Air Force), Australian Air Publication.

12. Adair, T. L. and Wehener, D. H. (1998) Automated Fluorescent Penetrant Inspection (FPI) System is Triple A. In: *IEEE Systems Readiness Technology Conference Proceedings*. Salt Lake City, Utah, USA, pp. 498-529.

13. *Non Destructive Testing General Procedures*. (1999) AAP 7002.043-36, Royal Australian Air Force, Australian Air Publication.

14. Packman, P. F., et al. (1968) *The Applicability of a Fracture Mechanics - Nondestructive Testing Design Criterion*. AFML-TR-68-32, Air Force Materials Laboratory, Ohio, USA.

15. Packman, P. F. (1973) Fracture Toughness / NDI Requirements for Aircraft Design. In: *TTCP Fracture Conference*. Vanderbilt University, Abington, Cambridge, June.

16. Rummel, W. D., et al. (1974) *The Detection of Fatigue Cracks by Nondestructive Testing Methods*. NASA CR 2369, National Aeronautics and Space Administration, Martin Marietta Aerospace, Denver, Colorado, USA.

17. Rummel, W. D. and Matzkanin, G. A. (1997) *Nondestructive Evaluation (NDE) Capabilities Data Book*. NTIAC-DB-97-02, Nondestructive Testing Information Analysis Center, Austin, Texas, USA.

18. Lewis, W. H., et al. (1978) *Reliability of Nondestructive Inspections - Final Report*. SA-ALC/MME 76-6-38-1, San Antonio Air Logistics Center, Kelly Air Force Base, USA.

19. *Military Specification Airplane Damage Tolerance Requirements*. (1974) MIL-A-83444, United States Air Force, USA.

20. Singh, R. (2000) *Three Decades of NDI Reliability Assessment*. Karto-3510-99-01, Karta Technologies Inc, San Antonio, Texas, USA.

21. Christner, B. K., Long, D. L. and Rummel, W. D. (1988) *NDE Detectability of Fatigue-Type Cracks in High-Strength Alloys: NDI Reliability Assessments*. MCR-33-1044, Martin Marietta Astronautics Group, Denver, Colorado, USA.

22. Fahr, A., et al. (1994) *NDI Techniques for Damage Tolerance-Based Life Predication of Aero-Engine Turbine Disks*. LTR-ST-1961, Structures, Materials and Propulsion Laboratory, Institute for Aerospace Research, Canada.

23. Fahr, A., et al. (1995) *POD Assessment of NDI Procedures Using a Round Robin Test*. AGARD-R-809, Advisory Group for Aerospace Research and Development, North Atlantic Treaty Organisation.

24. Bullock, M., Forsyth, D. and Fahr, A. (1994) *Statistical Functions and Computational Procedures for the POD Analysis of Hit/Miss NDI Data*. LTR-ST-1964, Institute for Aerospace Research, National Research Council, Ottawa, Canada.

25. Harding, C. A. and Hugo, G. R. (2003) Statistical Analysis of Probability of Detection Hit/Miss Data for Small Data Sets. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 22, pp. 1823-1844, American Institute of Physics, Melville, New York, USA.

26. Harding, C. A. (2008) *Methods for Assessment of Probability of Detection for Nondestructive Inspections*. PhD Thesis, The University of Melbourne, Victoria, Australia.

27. Forsyth, D. and Fahr, A. (1996) *The Sensitivity and Reliability of NDI Techniques for Gas Turbine Component Inspection and Life Prediction*. LTR-ST-2055, Institute for Aerospace Research, Canada.

28. Fahr, A. and Forsyth, D. (1998) POD Assessment Using Real Aircraft Engine Components. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 17, pp. 2005-2012, Plenum Press, New York, USA.

29. Fahr, A. and Forsyth, D. (1998) POD Measurement Using Actual Components. In: *Proceedings of SPIE: Nondestructive Evaluation of Aging Aircraft, Airports, and Aerospace Hardware II*. Vol. 3397, San Antonio, Texas, USA, 31 Mar - 2 Apr.

30. *FAA Center for Aviation Systems Reliability Program Summary*. (2002).

31. *Uncontained engine failure Delta Air Lines flight 1288 McDonnell Douglas MD-88 N927DA Pensacola, Florida July 6, 1996*. (1998) NTSB/AAR-98/01, National Transportation Safety Board.

32. Brasche, L. (2005) Engineering Assessment of Fluorescent Penetrant Inspection Program Overview. In: *Review of Progress in Quantitative Nondestructive Evaluation*. Brunswick, Maine, USA, 31 July - 5 August.

33. Brasche, L., Lopez, R. and Eisenmann, D. (2006) Characterization of Developer Application Methods Used in Fluorescent Penetrant Inspection. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 25, pp. 598-605, American Insitute of Physics, Melville, New York, USA.

34. Brasche, L. (2007) Engineering Studies of Fluorescent Penetrant Inspection. In: *Aging Aircraft Conference*. Palm Springs, California, USA, 16-19 April.

35. Brasche, L., et al. (2007) Update of FPI Engineering Studies. In: *Air Transport Association NDT Forum*. Orlando, Florida, USA, 27-30 August.

36. Eisenmann, D., Brasche, L. and Lopez, R. (2006) Preliminary Results of Cleaning Process for Lubricant Contamination. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 25, pp. 606-613, American Institute of Physics, Melville, New York, USA.

37. Brasche, L., et al. (2004) *Engineering Studies of Cleaning and Drying Processes in Preparation for Fluorescent Penetrant Inspection*. DOT/FAA/AR-03/73, Federal Aviation Administration, USA.

38. DiMambro, J., et al. (2007) Sonic Infrared (IR) Imaging and Fluorescent Penetrant Inspection Probability of Detection (POD) Comparison. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 26, pp. 463-470, AIP, Melville, New York, USA.

39. Mayton, D. (2007) *Nondestructive Evaluation Technology Initiatives II Delivery Order 0002: Whole Field Turbine Disk Inspection*. AFRL-RX-WP-TR-2008-4004, Science Applications International Coorporation, USA.

40. Brausch, J. C. and Tracy, N. A. (2001) *Effects of Compressive Stress on Fluorescent Penetrant Indications of Fatigue Cracks in Titanium*. AFRL-ML-WP-TR-2001-4139, Air Force Research Laboratory, Dayton, Ohio, USA.

41. Drury, C. G., et al. (2005) Fatigue Effects in Fluorescent Penetrant Inspection. In: *Aviation Maintenance Human Factors Program Review*. Office of the Chief Scientist for Human Factors, USA.

42. Lively, J. and Aljundi, T. L. (2003) Fluorescent Penetrant Inspection Probability of Detection Demonstrations Performed for Space Propulsion. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 22, pp. 1891-1898, AIP, Melville, New York, USA.

43. *Space systems - Fracture and damage control*. (2005) ISO 21347, International Standard.

44. Walpole, R. E. and Myers, R. H. (1993) *Probability and Statistics for Engineers and Scientists*. 5th ed, New York, Macmillan Publishing Company

45. Spencer, F. W. (1998) Identifying Sources of Variation for Reliability Analysis of Field Inspections. In: *RTO AVT Workshop on Airframe Inspection Reliability Under Field/Depot Conditions*. NATO RTO-MP-10, Brussels, Belgium, 13-14 May.

46. *Task 07/101 - AIR NDE Research Technical Aspects*. Department of Defence Corporate File 2007/1108056/2, Originating workgroup: DSTO Melbourne, Australia.

Appendix A: Summary of Statistical Inferences on a_{90}

Table A. 1 in this appendix provides a summary of each of the four main inference statistics used in Section 4.1. These are separated into:

- statistics relating to the average POD performance across all implementations of a given NDT method or process; and
- statistics that consider that the variability in POD performance between different implementations of an NDT method or process (e.g. between different facilities) and anticipate the worst performance likely to be encountered.

Table A. 1 also includes a discussion of the use of POD curves fitted to raw hit/miss data pooled across a variety of implementations as presented in Section 4.2.

Table A. 1 Summary of statistical inferences on a_{90}

Statistic	Explanation	Confidence that value will capture range of performance of a method on ADF aircraft
A. Statistics relating to the average performance across all implementations of an NDT method:		
$\hat{\mu}_{a_{90}}$	This value represents the best statistical estimate of the median of the measured values of a_{90} for the distribution of different implementations captured in the data set.[14] Assuming the individual measurements of a_{90} are unbiased, this may be regarded as a point estimate of the *average* a_{90} achieved across the range of implementations represented by the POD trial data. The $\hat{\mu}_{a_{90}}$ statistic does not allow for: (i) variability in the POD performance between facilities, nor (ii) the randomness inherent in the statistical measurement of POD. There is a 50% risk (approximately) that the actual a_{90} achieved for any real implementation will exceed $\hat{\mu}_{a_{90}}$. Hence, if this parameter were to be used as a_{NDI} there would be a 50% risk that this a_{NDI} would not be achieved in practice.	VERY LOW
Confidence interval on $\mu_{a_{90}}$	The confidence interval defines a range within which it likely that the true median of a_{90} will lie, taking into account the randomness inherent in measuring a small sample of a_{90} values. For example, for a 95% confidence interval, we can be 95% confident that the upper bound of the interval will not be exceeded by the true average (median) a_{90} for the NDT method. However, the confidence interval on the median considers only the average POD performance of a method and does not consider variability in the POD reliability performance between facilities. Thus, we cannot be confident that the interval will necessarily capture the range of a_{90} values achieved across individual implementations.	LOW

[14] The median is the central value above and below which 50% of a distribution lies and is a useful alternative to the traditional arithmetic mean as a measure of the average. For the log-normal distribution fitted to the a_{90} values in Figures 7 and 8, the fitted location parameter, $\hat{\mu}_{a_{90}}$, is actually the median a_{90} value, rather than the mean.

Table A. 1 Summary of statistical inferences on a_{90} (continued)

Statistic	Explanation	Confidence that value will capture range of performance of a method on ADF aircraft
B. Statistics that anticipate the variation in performance likely to be encountered across a range of implementations of an NDT method:		
y% prediction limit on a_{90} (one-sided upper limit)	This value gives the 'best estimate' of the a_{90} value expected to be achieved or bettered by y% of actual implementations, based on the available trials data. This value explicitly accounts for the *observed* variability in the available data between a_{90} measurements for the different implementations of the NDT method. It does not account for the sampling variability inherent in the selection of the trial data from the range of implementations used in aircraft maintenance. Risks in using this value for a_{NDI}: (i) the variability between implementations captured in the trial data may not adequately represent the variability present in Australian implementations used on ADF aircraft. (ii) The performance of Australian implementations could be, on average, worse than that captured in the trial data.	MODERATE
y% tolerance limit on a_{90} with z% confidence (one-sided upper limit)	This value gives an upper bound on the a_{90} value expected to be achieved or bettered by y% of actual implementations with z% statistical confidence based on the available data. This value allows for both the variability in performance between different implementations and the sampling variability inherent in the selection of the trial data from the range of implementations used in aircraft maintenance. This value represents the most rigorous and conservative choice for a_{NDI}. Risks in using this value for a_{NDI}: There is a high degree of confidence that this value will not be smaller than the actual a_{90} achieved by any particular implementation. As such, it is highly conservative and could be unduly pessimistic about the actual performance (particularly if based on a small data set). As this number will be relatively large, its use may result in: (i) dramatically reduced inspection intervals, and/or (ii) reluctance by engineering staff to accept the use of the method, even when it is the most appropriate NDT method.	HIGH

Table A. 1 Summary of statistical inferences on a_{90} (continued)

Statistic	Explanation	Confidence that value will capture range of performance of a method on ADF aircraft
C. Statistics that relate to the demonstrated POD as a function of defect size for data pooled across all the different organisations for which POD data was obtained or published:		
$a_{90/95}$ computed for pooled data set (Refer to section 4.2 for example[15])	This value is the traditional $a_{90/95}$ defect size demonstrated to be detected with 90% POD and 95% statistical confidence based on a set of inspection results (as specified in airworthiness standards such as JSSG-2006 [2]). Refer to [1] for a full explanation of $a_{90/95}$. Pooling the data into a single data set assumes that there is a good level of consistency in the POD achieved between the different implementations. The $a_{90/95}$ value represents a confidence limit on the notional *average* POD achieved across all implementations captured in the trial data. The $a_{90/95}$ confidence limit provides no assessment of the potential variability between implementations. If there is significant variability between implementations, the $a_{90/95}$ value provides no specific level of confidence that it will be a conservative assessment of the a_{90} achieved for any individual implementation. Risks in using this value for a_{NDI}: If there is significant variability between the performances of the different implementations for which data are available, then pooling these data is not valid. Consequently, use of the pooled $a_{90/95}$ value alone without consideration of the other statistics carries an unknown risk that some implementations of the NDT method would fail to achieve this POD performance.	MODERATE

[15] The $a_{90/95}$ values reported in section 4.2 were obtained using the available hit/miss data for NRC IAR POD trials, pooling the results from eight different organisations (and implementations of LPT) into a single combined data set.

Appendix B: Inconsistencies with NRC IAR Data from NDE Capabilities Data Book

B.1. Inconsistencies within the NRC IAR Reports

For the NATO round-robin trials [22], the summary of inspection results in the body of the report (Table II) lists different size data sets than are contained in the raw hit/miss data in Appendix B of the same report, as outlined in Table B. 1. These differences cannot be due to rejection of small cracks because there are some extra data points and some missing.

For the Canadian round-robin trials [27], a similar comparison indicates that the summary statistics in Appendix C and Table I of reference [27] agree when rejection of hits < 0.3 mm is taken into account.

B.2. Inconsistencies within the NDE Capabilities Data Book

The chart for organisation 2, 4th stage data lists n = 216 and h = 62, whereas the spreadsheet actually contains n = 255 and h = 61.

B.3. Inconsistencies between NDE Capabilities Data Book Spreadsheets and Hit/Hiss Data in NRC IAR Appendices

For the Canadian round-robin trials [27], there were no discrepancies found between Appendix C of [27] and the NDE Capabilities Data Book spreadsheets.

For the NATO round-robin trials [22], the following discrepancies have been found:

- For organisation 1, the Data Book spreadsheets do not provide disk ID and hole numbers for 166 misses and 42 hits. These data are almost certainly for disks H1 to M2, which then gives correct totals for Organisation 1.
- Data for organisation 2, disk H1, (0 hits, 6 misses) and organisation 3, disks I1, J1, K1, L1 (total 0 hits, 102) misses are absent from the Data Book spreadsheets. There are no hits in these missing data but identical results from other organisations on these disks were included in the spreadsheets. There is no explanation as to why these data are not present in NDE capabilities data book spreadsheets.

Table B. 1 Differences in summaries of inspection results within reference [22]

	Table II of [27]			**Appendix B of [27]**			**Table II lists:**
	hits	**misses**	**total**	**hits**	**misses**	**total**	
Org 1	73	212	285	73	223	296	11 fewer misses
Org 2	101	303	404	101	309	410	6 fewer misses
Org 3	52	155	207	52	138	190	17 extra misses
Org 4	79	206	285	79	217	296	11 fewer misses
Org 6	15	119	134	4	53	57	11 extra hits and 66 extra misses
		Total	1315		Total	1249	Total of 66 fewer data points in Table II than Appendix B

B.4. Other issues with data sets used to generate published POD curves

As mentioned in Section 3.3, the analysis performed by NRC IAR on data from organisations A and D excluded hits for defects of size < 0.3 mm [27].

In the NDE Capabilities Data Book, the analysis of data from reference [27] deliberately excluded some data, as described in Table B. 2. It is surprising that the cut-off between which data are included and which are excluded does not appear to be purely on the basis of defect size. Also, data from other organisations is apparently similar in this size range but has not been excluded from analysis.

Table B. 2 Data deliberately excluded from NDE Capabilities Data Book analysis

	No. data points excluded	**Sizes excluded**	**Comment from NDE Capabilities Data Book spreadsheet**
Org C	40	≤ 0.1 mm (Some defects of size 0.1 mm are included and some are excluded.)	"Data not used. Data are divergent and hence not rigorous." [17]
Org D	20	≤ 0.07 mm (Some defects of size 0.07 mm are included and some are excluded.)	"Data not used. Data are divergent and hence not rigorous when crack distribution is not balanced. (Author noted poor fit of data.)" [17]

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Review of Literature on Probability of Detection for Liquid Penetrant Nondestructive Testing

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

C.A. Harding and G.R. Hugo

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend Victoria 3207 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2623	AR-015-142	Technical Report	November 2011

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2009/1047084/1	AIR 07/101	RAAF DGTA	40	46

13. DOWNGRADING/DELIMITING INSTRUCTIONS

To be reviewed three years after date of publication

14. RELEASE AUTHORITY

Chief, Maritime Platforms Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.shtml

Nondestructive testing, Structural integrity, Reliability, Probability

19. ABSTRACT

A review of the published literature on the reliability of liquid penetrant testing (LPT) identified twelve major probability of detection (POD) studies conducted between 1968 and 2009. Based on these studies, significant variability in performance is inferred between different implementations of the post-emulsifiable LPT process. This report presents statistical inferences for the defect size expected to be detected with 90% POD by most implementations of post-emulsifiable LPT, based on the published data. The reliably-detectable defect size of aNDI = 3 mm currently specified for the Royal Australian Air Force general procedure for post-emulsifiable LPT is consistent with estimates of the average performance of LPT demonstrated in the literature.